# Week 3 Assignment

By: Zach Adair

MSDS 662: Exploratory Data Analysis

Regis University

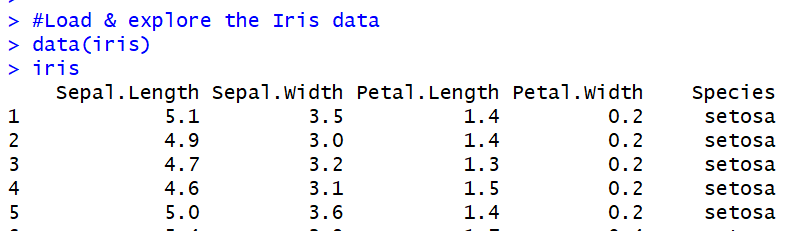
9/12/2019

## Introduction

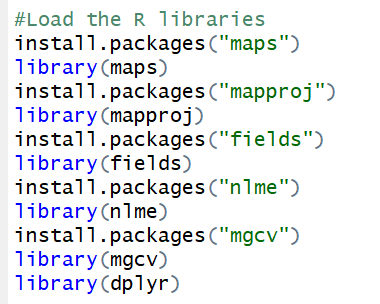
This week in class, we have gone over the exploratory graphs and plotting systems. In RStudio, this is a great way to explore the data by providing a visual representation of what the data is doing to find out more about its distribution, how its trending or what kind of volumes we are seeing based on specific factors or characteristics. In this exercise, I will be using the Iris dataset and creating charts and graphs which will visualize this data and help gain a deeper understanding of this data and what it is doing.

## Load the Data

Before analyzing the data we need to load our data and the library packages we will need.

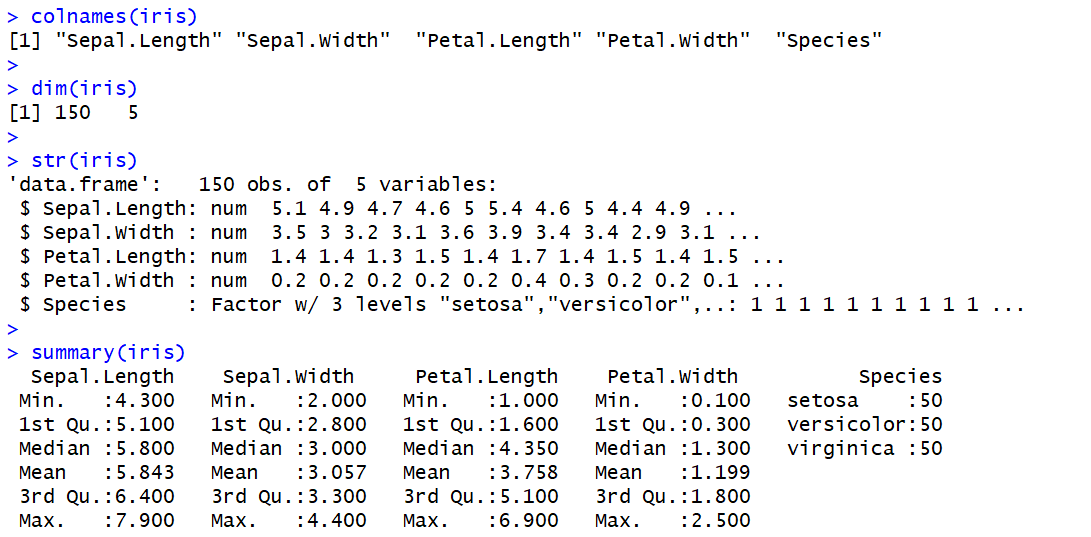


The libraries I loaded were:



## Initial Data Exploration

Now that I have the data and RStudio packages loaded in, its time to do a little data exploration before I begin with the visualizations.



From our initial data exploration, we see that the data has 150 records and 5 variables. Also, we see that 4/5 variables are numeric and only one is factorial and that is species. When we look at the summary of our data, we see the summary stats on our sepal and petals lengths and widths and we also see an even distribution between the species, with the 3 species each having 50 entries.

## Visualizations of the Iris Data

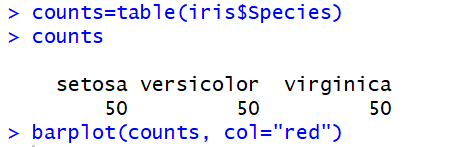
There are many visualizations that can show what is going on within a dataset in RStudio. With the Iris dataset we will go through a few of them and explain what we are seeing with each one.

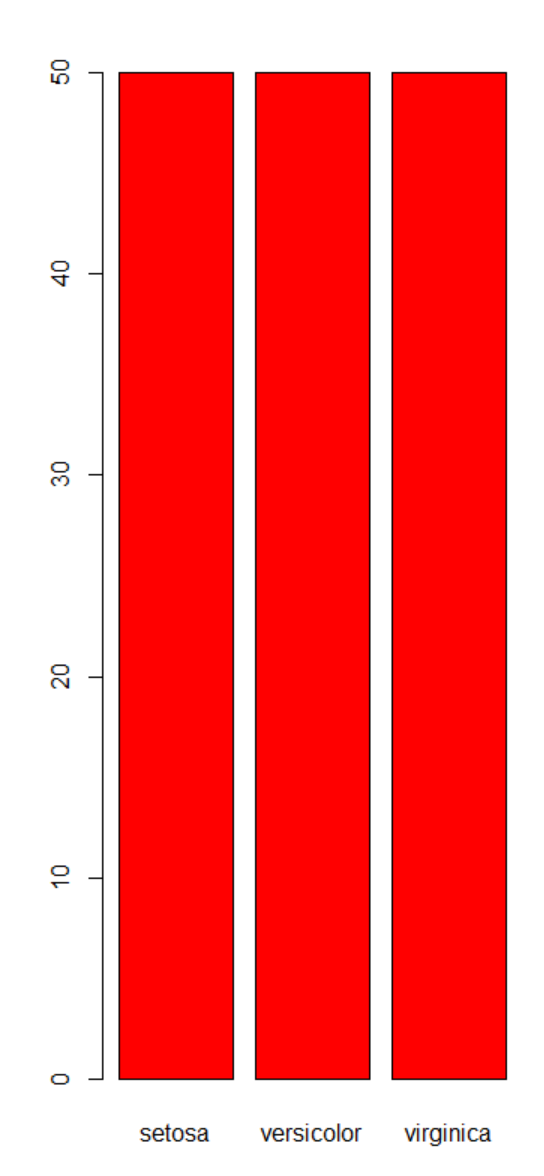
### Bar Chart

Bar charts are an effective way to show the distribution of multiple factors to see how they are spread throughout a dataset.



This initial attempt at a bar chart was unsuccessful, mainly because of how I set it up and putting the variable of setosa as a color when it is obviously not. But what we can do to insure we see a bar chart of the species is by creating a table of the iris species.





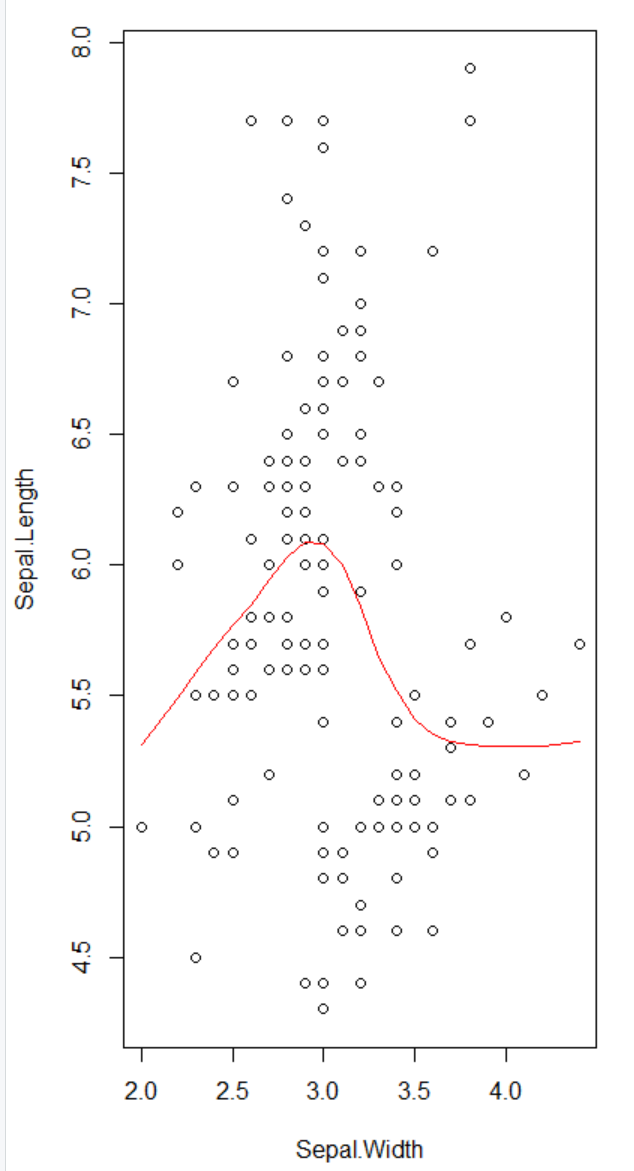
As you can see above a bar chart can be created from the table which did a count of the different Iris species. They all have 50 records so that is why the bar chart has 3 equal bars.

### Line Graphs

Line graphs do a great representation of showing how data moves over time, or it can show if there is a possible relationship between two variables. With there being no variable in accordance to time in the Iris dataset we can only use line charts as a tool for regression and try to map relationships between variables.

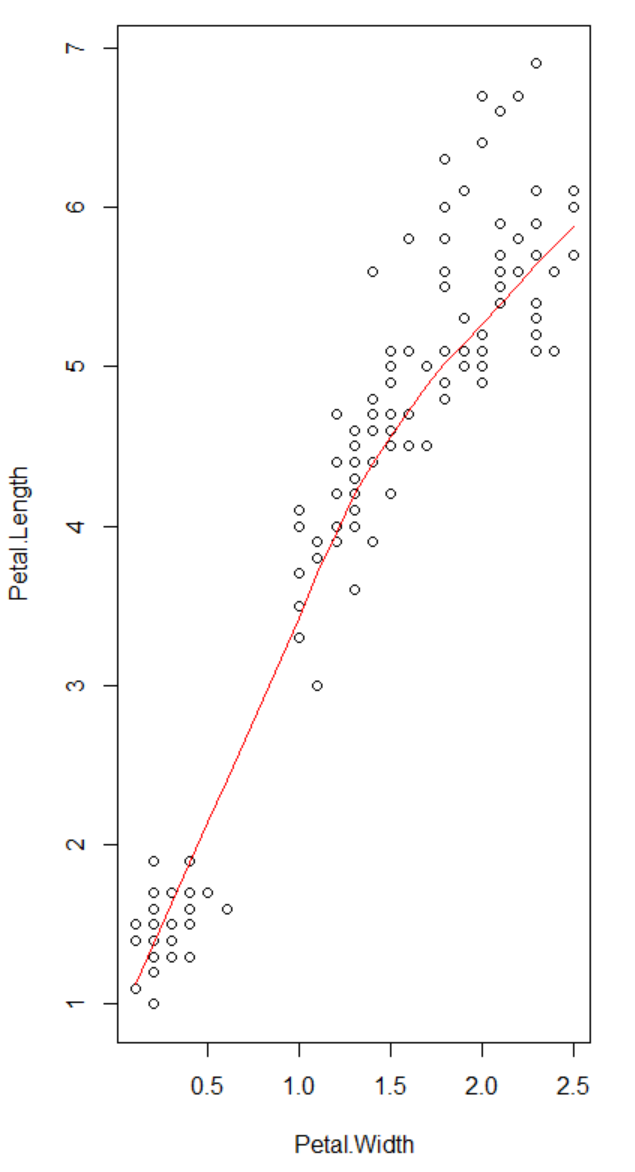
**Sepal Length and Sepal Width**





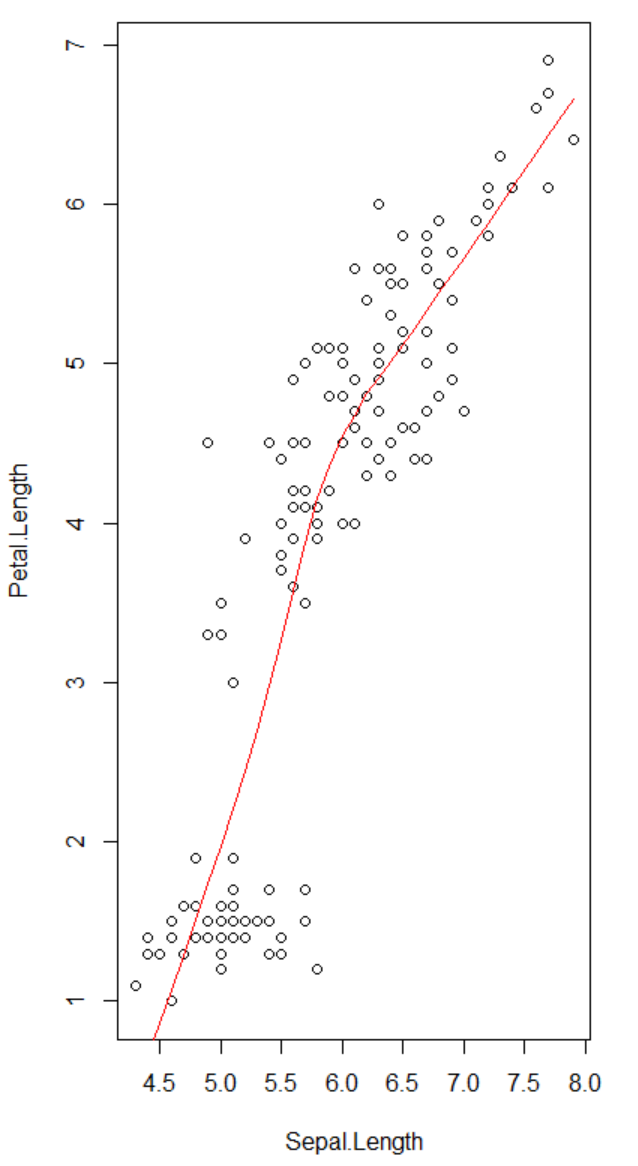
**Petal Length vs Petal Width**





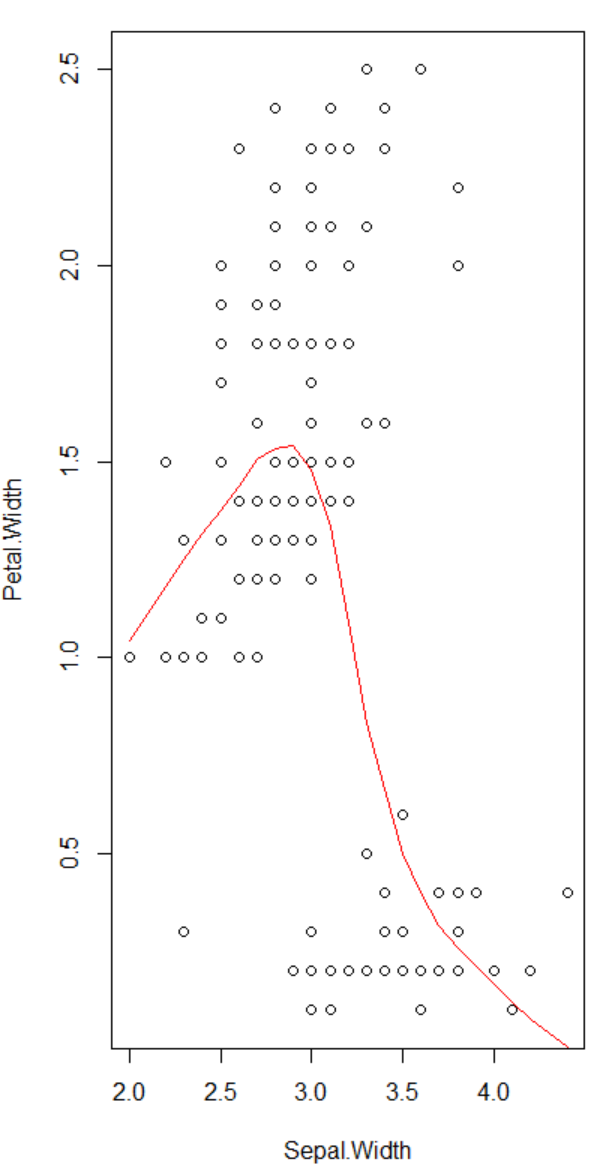
**Petal Length vs Sepal Length**





**Petal Width vs Sepal Width**



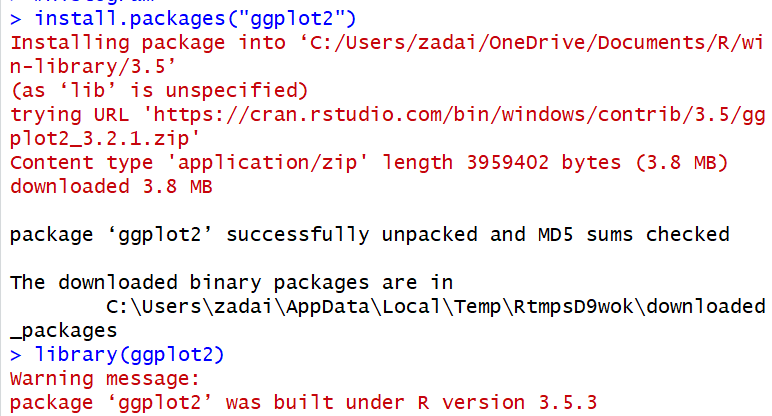


From the four plots we have in our data, some pretty distinct charts based on the spread of the data points and how the line goes with the data points. In the first chart we get a bit of a parabolic shape with our line with the line moving negatively and positively depending on its spot on the chart, but doesn’t really give us any indication of a relationship between Sepal length and width. The next two, which deal with Petal length and width as well as petal length and sepal length appear to have a positive relationship because of how the line grows as the variables grow. Then when comparing the widths of petals and sepals there doesn’t appear to be much of a relationship based on the line with the distribution more spread out and the line beginning with a positive gain but ends going in a negative slope.

The last visualization I will display of the Iris data will be a histogram.

### Histogram

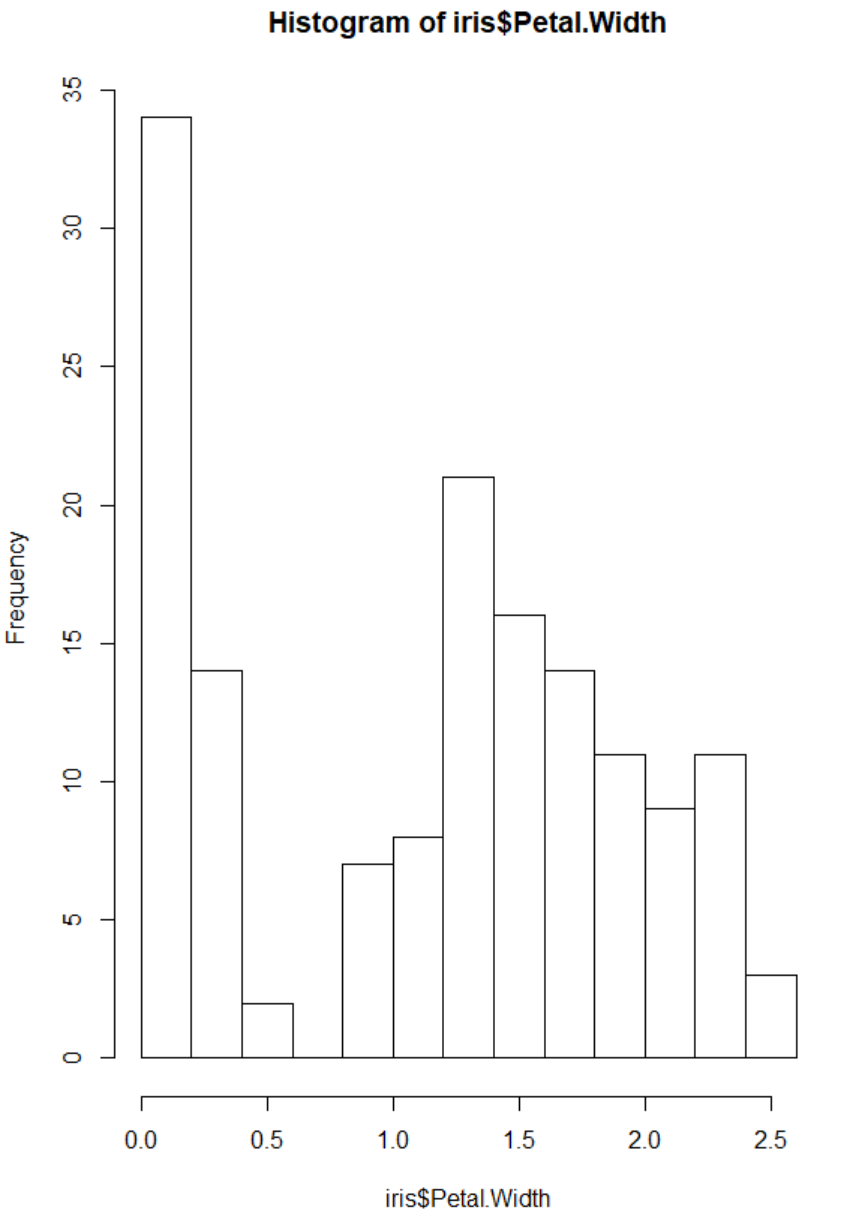
Before getting started I will also need to bring in the ggplot2 library.



Now that the data is loaded, I will create a histogram for each of the numeric variables.

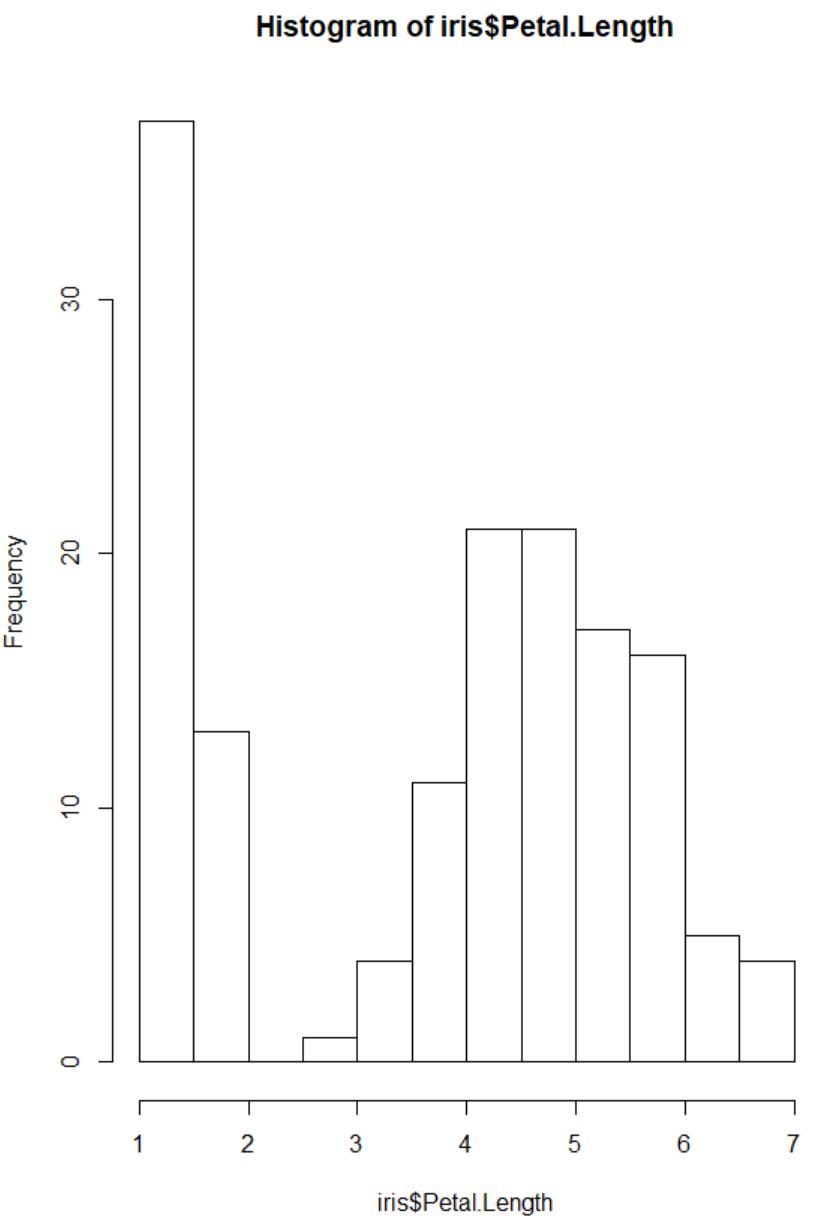
**Petal Width**





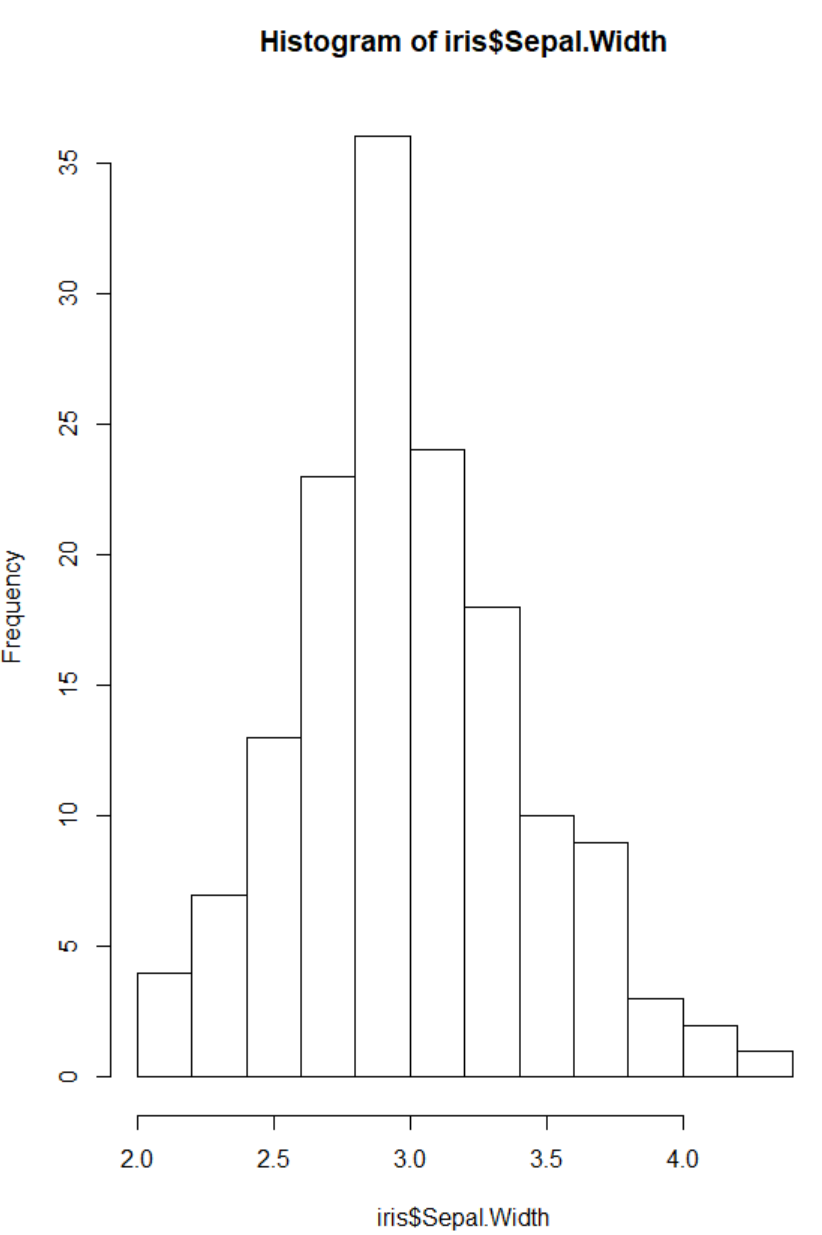
**Petal Length**





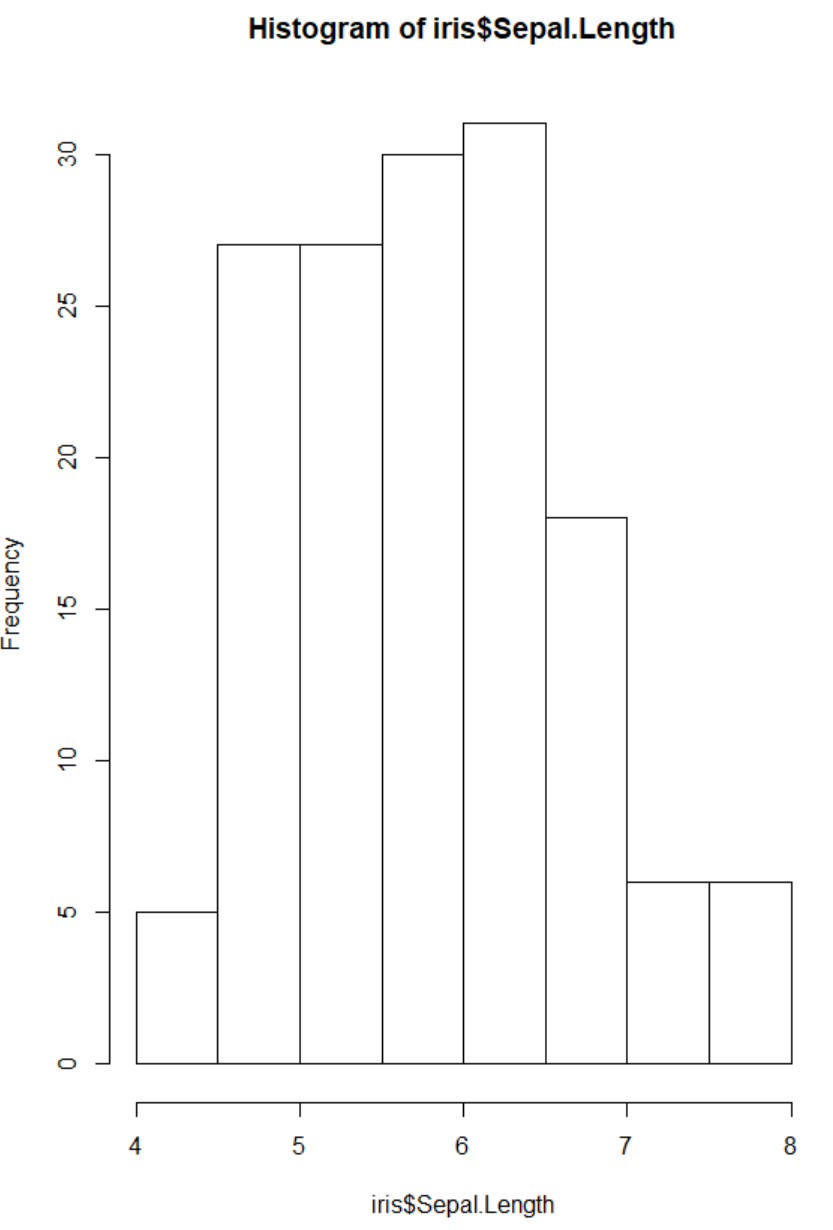
**Sepal Width**





**Sepal Length**





The benefit histograms give are they help with understanding the distribution of the data. Typically we are looking for our data to have a normal distribution, this is where our highest bins are the ones in the center and then the frequency of the bins on the outside are smaller. The normal distribution is also called a bell shaped curve for how it is distributed. From our variables in the Iris data it appears that we only have one distribution that is normal, the sepal width. The sepal length is somewhat normal, but the distribution on the left is much greater per bin than on the left side. Then both the petal length and width have skewed distributions to the left and are very identical, but if we remember when we compared the distribution with a scatter plot and line chart the way our line moved and our data points were plotted we could have suspected that the distribution of each variable would be similar.

## Reflection

Visualizing the data is another important step in the Exploratory Data Analysis process. It is a great opportunity to show any sort of distribution or trend in the data that is hard to distinguish within a typical data table. This is a step that should be implemented right after the initial exploration, after finding out what are the characteristics of the data sets rows and columns. Doing this before any data modeling can help an analyst understand what a data model or machine learning model outcome could be before doing that with the data.